

Bayesian optimization under mixed constraints with a slack-variable augmented Lagrangian

Victor Picheny* Robert B. Gramacy[†] Stefan M. Wild[‡]
Sébastien Le Digabel[§]

June 1, 2016

Abstract

An augmented Lagrangian (AL) can convert a constrained optimization problem into a sequence of simpler (e.g., unconstrained) problems, which are then usually solved with local solvers. Recently, surrogate-based Bayesian optimization (BO) sub-solvers have been successfully deployed in the AL framework for a more global search in the presence of inequality constraints; however, a drawback was that expected improvement (EI) evaluations relied on Monte Carlo. Here we introduce an alternative slack variable AL, and show that in this formulation the EI may be evaluated with library routines. The slack variables furthermore facilitate equality as well as inequality constraints, and mixtures thereof. We show how our new slack “ALBO” compares favorably to the original. Its superiority over conventional alternatives is reinforced on several mixed constraint examples.

1 Introduction

Bayesian optimization (BO), as applied to so-called *blackbox* objectives, is a modernization of 1970-80s statistical response surface methodology for sequential design (Mockus et al., 1978; Box and Draper, 1987; Mockus, 1989, 1994). In BO, nonparametric (Gaussian) processes (GPs, Rasmussen and Williams, 2006) provide flexible response surface fits. Sequential design decisions, so-called *acquisitions*, judiciously balance exploration and exploitation in search for global optima. For a review of GP surrogate modeling and optimization in the context of computer experiments, see Santner et al. (2003), Booker et al. (1999) and Bingham et al. (2014). For a machine learning perspective, see Brochu et al. (2010), and Boyle (2007). Until recently, most works in these literatures have focused on unconstrained optimization.

*Institut National de Recherche Agronomique–Centre de Toulouse, France

[†]Corresponding author: The University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago IL, 60605; rbgramacy@chicagobooth.edu

[‡]Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439

[§]GERAD and Département de Mathématiques et Génie Industriel, École Polytechnique de Montréal, Montréal, QC H3C 3A7, Canada

Many interesting problems contain constraints, typically specified as equalities or inequalities:

$$\min_x \{f(x) : g(x) \leq 0, h(x) = 0, x \in \mathcal{B}\}, \quad (1)$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a scalar-valued objective function, and $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$ and $h : \mathbb{R}^d \rightarrow \mathbb{R}^p$ are vector-valued constraint functions taken componentwise (i.e., $g_j(x) \leq 0$, $j = 1, \dots, m$; $h_k(x) = 0$, and $k = 1, \dots, p$). We differentiate these constraints from $\mathcal{B} \subset \mathbb{R}^d$, which is a known (not blackbox) set, typically containing bound constraints; here we assume that \mathcal{B} is a bounded hyperrectangle.¹

The typical setup treats f , g , and h as a “joint” blackbox, meaning that providing x to a single computer code reveals $f(x)$, $g(x)$, and $h(x)$ simultaneously, often at great computational expense. A common special case treats $f(x)$ as known (e.g., linear); however, the problem is still hard when $g(x) \leq 0$ defines a nonconvex valid region. A tacit goal in this setup is to minimize the number of blackbox runs required to solve the problem, and a common meter of progress tracks the best valid value of the objective (up to a relaxing threshold on the equality constraints h) at increasing computational budgets (# of evaluations, n).

Not many algorithms target global solutions to this general, constrained blackbox optimization problem. Statistical methods are acutely few. We know of no methods from the BO literature natively accommodating equality constraints, let alone mixed (equality and inequality) ones. Schonlau et al. (1998) describe how their expected improvement (EI) heuristic can be extended to multiple inequality constraints by multiplying by an estimated probability of constraint satisfaction. Here, we call this *expected feasible improvement (EFI)*. EFI has recently been revisited by several authors (Snoek et al., 2012; Gelbart et al., 2014; Gardner et al., 2014). However, the technique has pathological behavior in otherwise idealized setups (Gramacy et al., 2016b), which is related to a so-called “decoupled” pathology (Gelbart et al., 2014). Some recent information-theoretic alternatives have shown promise in the inequality constrained setting (Hernández-Lobato et al., 2015; Picheny, 2014); tailored approaches have shown promise in more specialized constrained optimized setups (Gramacy and Lee, 2011; Williams et al., 2010; Parr et al., 2012).

We remark that any problem with equality constraints can be “transformed” to inequality constraints only, by applying $h(x) \leq 0$ and $h(x) \geq 0$ simultaneously. However, the effect of such a reformulation is rather uncertain. It puts double-weight on equalities and violates certain regularity (i.e., constraint qualification (Nocedal and Wright, 2006)) conditions. Numerical issues have been reported in empirical evaluations (Sasena, 2002). In our own empirical work [Section 4] we find unfavorable performance.

In this paper we show how a recent technique (Gramacy et al., 2016a) for BO under inequality constraints is naturally enhanced to handle equality constraints, and therefore mixed ones too. The method involves converting inequality constrained problems into a

¹The presumption throughout is that a solution of (1) exists so that the feasible region $\{x \in \mathbb{R}^d : g(x) \leq 0, h(x) = 0\} \cap \mathcal{B}$ is nonempty; however, the algorithms we describe can provide reasonable approximate solutions when the feasible region is empty, so long as the modeling assumptions (e.g., smoothness and stationarity) on f , g , and h , and constraint qualifications (see, e.g., (Nocedal and Wright, 2006)) are not egregiously violated.

sequence of simpler subproblems via the augmented Lagrangian (AL, Bertsekas, 1982)). AL-based solvers can, under certain regularity conditions, be shown to converge to locally optimal solutions that satisfy the constraints, so long as the sub-solver converges to local solutions. By deploying modern BO on the subproblems, as opposed to the usual local solvers, the resulting meta-optimizer is able to find better, less local solutions with fewer evaluations of the expensive blackbox, compared to several classical and statistical alternatives. Here we dub that method ALBO.

To extend ALBO to equality constraints, we suggest the opposite transformation to the one described above: we convert inequality constraints into equalities by introducing *slack variables*. In the context of earlier work with the AL, via conventional solvers, this is rather textbook (Nocedal and Wright, 2006, Ch. 17). Handling the inequalities in this way leads naturally to solutions for mixed constraints and, more importantly, dramatically improves the original *inequality-only* version. In the original (non-slack) ALBO setup, the density and distribution of an important composite random predictive quantity is not known in closed form. Except in a few particular cases (Picheny et al., 2016), calculating EI and related quantities under the AL required Monte Carlo integration, which means that acquisition function evaluations are computationally expensive, noisy, or both. A reformulated slack-AL version emits a composite that has a known distribution, a so-called weighted non-central Chi-square (WNCS) distribution. We show that, in that setting, EI calculations involve a simple 1-d integral via ordinary quadrature. Adding slack variables increases the input dimension of the optimization subproblems, but only artificially so. The effects of expansion can be mitigated through optimal default settings, which we provide.

The remainder of the paper is organized as follows. Section 2 outlines the components germane to the ALBO approach: AL, Bayesian surrogate modeling, and acquisition via EI. Section 3 contains the bulk of our methodological contribution, reformulating the AL with slack variables and showing how the EI metric may be calculated in closed form, along with optimal default slack settings, and open-source software. Implementation details are provided in our Appendix. Section 4 provides empirical comparisons, and Section 5 concludes.

2 A review of relevant concepts: EI and AL

2.1 Expected improvement

The canonical acquisition function in BO is *expected improvement* (EI, Jones et al., 1998). Consider a surrogate $f^n(x)$, trained on n pairs $(x_i, y_i = f(x_i))$ emitting Gaussian predictive equations with mean $\mu^n(x)$ and standard deviation $\sigma^n(x)$. Define $f_{\min}^n = \min_{i=1,\dots,n} y_i$, the smallest y -value seen so far, and let $I(x) = \max\{0, f_{\min}^n - Y(x)\}$ be the *improvement* at x . $I(x)$ is largest when $Y(x) \sim f^n(x)$ has substantial distribution below f_{\min}^n . The expectation of $I(x)$ over $Y(x)$ has a convenient closed form, revealing balance between exploitation ($\mu^n(x)$ under f_{\min}^n) and exploration (large $\sigma^n(x)$):

$$\mathbb{E}\{I(x)\} = (f_{\min}^n - \mu^n(x))\Phi\left(\frac{f_{\min}^n - \mu^n(x)}{\sigma^n(x)}\right) + \sigma^n(x)\phi\left(\frac{f_{\min}^n - \mu^n(x)}{\sigma^n(x)}\right), \quad (2)$$

where Φ (ϕ) is the standard normal cdf (pdf). Accurate, approximately Gaussian predictive equations are provided by many statistical models (GPs, Rasmussen and Williams, 2006).

When the predictive equations are not Gaussian, Monte Carlo schemes—sampling $Y(x)$ variates from $f^n(x)$ and averaging the resulting $I(x)$ values—offer a suitable, if computationally intensive, alternative. Yet most of the theory applies only to the Gaussian case. Under certain regularity conditions (e.g., Bull, 2011), one can show that sequential designs derived by repeating the process, generating $(n+1)$ -sized designs from n -sized ones, contain a point x_i whose associated $y_i = f(x_i)$ is arbitrarily close to the true minimum objective value, that is, the algorithm converges.

2.2 Augmented Lagrangian framework

Although several authors have suggested extensions to EI for constraints, the BO literature has primarily focused on unconstrained problems. The range of constrained BO options was recently extended by borrowing an apparatus from the mathematical optimization literature, the *augmented Lagrangian*, allowing unconstrained methods to be adapted to constrained problems. The AL, as a device for solving problems with inequality constraints (no $h(x)$ in Eq. (1)), may be defined as

$$L_A(x; \lambda, \rho) = f(x) + \lambda^\top g(x) + \frac{1}{2\rho} \sum_{j=1}^m \max\{0, g_j(x)\}^2, \quad (3)$$

where $\rho > 0$ is a *penalty parameter* on constraint violation and $\lambda \in \mathbb{R}_+^m$ serves as a *Lagrange multiplier*. AL methods are iterative, involving a particular sequence of $(x; \lambda, \rho)$. Given the current values ρ^{k-1} and λ^{k-1} , one approximately solves the subproblem

$$\min_x \{L_A(x; \lambda^{k-1}, \rho^{k-1}) : x \in \mathcal{B}\}, \quad (4)$$

via a conventional (bound-constrained) solver. The parameters (λ, ρ) are updated depending on the nature of the solution found, and the process repeats. The particulars in our setup are provided in Alg. 1; for more details see (Nocedal and Wright, 2006, Ch. 17). Local convergence is guaranteed under relatively mild conditions involving the choice of subroutine solving (4). Loosely, all that is required is that the solver “makes progress” on the subproblem. In contexts where termination depends more upon computational budget than on a measure of convergence, as in many BO problems, that added flexibility is welcome.

However, the AL does not typically enjoy global scope. The local minima found by the method are sensitive to initialization—of starting choices for (λ^0, ρ^0) or x^0 ; local searches in iteration k are usually started from x^{k-1} . However, this dependence is broken when statistical surrogates drive search for solutions to the subproblems. Independently fit GP surrogates, $f^n(x)$ for the objective and $g^n(x) = (g_1^n(x), \dots, g_m^n(x))$ for the constraints, yield predictive distributions for $Y_f^n(x)$ and $Y_g^n(x) = (Y_{g_1}^n(x), \dots, Y_{g_m}^n(x))$. Dropping the n superscripts, the AL composite random variable

$$Y(x) = Y_f(x) + \lambda^\top Y_g(x) + \frac{1}{2\rho} \sum_{j=1}^m \max\{0, Y_{g_j}(x)\}^2 \quad (5)$$

Require: $\lambda^0 \geq 0, \rho^0 > 0$

- 1: **for** $k = 1, 2, \dots$ **do**
- 2: Let x^k (approximately) solve (4)
- 3: Set $\lambda_j^k = \max\left(0, \lambda_j^{k-1} + \frac{1}{\rho^{k-1}} g_j(x^k)\right), j = 1, \dots, m$
- 4: If $g(x^k) \leq 0$, set $\rho^k = \rho^{k-1}$; otherwise, set $\rho^k = \frac{1}{2}\rho^{k-1}$
- 5: **end for**

Algorithm 1: Basic augmented Lagrangian method.

can serve as a surrogate for (3); however, it is difficult to deduce the full distribution of this variable from the components of Y_f and Y_g , even when those are independently Gaussian. While its mean is available in closed form, EI requires Monte Carlo.

3 A novel formulation involving slack variables

An equivalent formulation of (1) involves introducing slack variables, s_j , for $j = 1, \dots, m$ (i.e., one for each inequality constraint $g_j(x)$), and converting the mixed constraint problem (1) into one with only equality constraints (plus bound constraints for s_j):

$$g_j(x) - s_j = 0, \quad s_j \in \mathbb{R}_+, \quad \text{for } j = 1, \dots, m. \quad (6)$$

The input dimension of the problem is increased from d to $d + m$ with the introduction of the m (positive) slack “inputs” s .

Reducing a mixed constraint problem to one involving only equality and bound constraints is valuable insofar as one has good solvers for those problems. Indeed, the AL method is an attractive option here, but some adaptation is required. Suppose, for the moment, that the original problem (1) has no equality constraints (i.e., $p = 0$). In this case, a slack variable-based AL method is readily available—as an alternative to the one described in Section 2.2. Although we frame it as an “alternative”, some in the mathematical optimization community would describe this as the standard version (see, e.g., Nocedal and Wright, 2006, Ch. 17). The AL for this setup is given as follows:

$$L_A(x, s; \lambda_g, \rho) = f(x) + \lambda^\top (g(x) + s) + \frac{1}{2\rho} \sum_{j=1}^m (g_j(x) + s_j)^2. \quad (7)$$

This formulation is more convenient than (3) because the “max” is missing, but the extra slack variables mean solving a higher $(d + m)$ dimensional subproblem compared to (4). In Section 3.3 we address this issue, and further detail adjustments to Alg. 1 for the slack variable case.

The AL can be expanded to handle equality (and thereby mixed constraints) as follows:

$$L_A(x, s; \lambda_g, \lambda_h, \rho) = f(x) + \lambda_g^\top (g(x) + s) + \lambda_h^\top h(x) + \frac{1}{2\rho} \left[\sum_{j=1}^m (g_j(x) + s_j)^2 + \sum_{k=1}^p h_k(x)^2 \right]. \quad (8)$$

Defining $c(x) := [g(x)^\top, h(x)^\top]^\top$, $\lambda := [\lambda_g^\top, \lambda_h^\top]^\top$, and enlarging the dimension of s with the understanding that $s_{m+1} = \dots = s_{m+p} = 0$, leads to a streamlined AL for mixed constraints

$$L_A(x, s; \lambda, \rho) = f(x) + \lambda^\top (c(x) + s) + \frac{1}{2\rho} \sum_{j=1}^{m+p} (c_j(x) + s_j)^2, \quad (9)$$

with $\lambda \in \mathbb{R}^{m+p}$. A non-slack AL formulation (3) may be written analogously as

$$L_A(x; \lambda_g, \lambda_h, \rho) = f(x) + \lambda_g^\top g(x) + \lambda_h^\top h(x) + \frac{1}{2\rho} \left[\sum_{j=1}^m \max\{0, g_j(x)\}^2 + \sum_{k=1}^p h_k(x)^2 \right],$$

with $\lambda_g \in \mathbb{R}_+^m$ and $\lambda_h \in \mathbb{R}^p$. Eq. (9), by contrast, is easier to work with because it is a smooth quadratic in the objective (f) and constraints (c). In what follows, we show that (9) facilitates calculation of important quantities like EI, in the GP-based BO framework, via a library routine. So slack variables not only facilitate mixed constraints in a unified framework, they also lead to a more efficient handling of the original inequality (only) constrained problem.

3.1 Distribution of the slack-AL composite

If Y_f and $Y_{c_1}, \dots, Y_{c_{m+p}}$ represent random predictive variables from $m + p + 1$ surrogates fitted to n realized objective and constraint evaluations, then the analogous slack-AL random variable is

$$Y(x, s) = Y_f(x) + \sum_{j=1}^{m+p} \lambda_j (Y_{c_j}(x) + s_j) + \frac{1}{2\rho} \sum_{j=1}^{m+p} (Y_{c_j}(x) + s_j)^2. \quad (10)$$

As in the original AL formulation, the mean of this RV has a simple closed form in terms of the means and variances of the surrogates, regardless of the form of predictive distribution. In the conditionally Gaussian case, we can derive the full distribution of the slack-AL variate (10) in closed form. Toward that aim, we re-develop the composite Y as follows:

$$\begin{aligned} Y(x, s) &= Y_f(x) + \sum_{j=1}^{m+p} \lambda_j s_j + \frac{1}{2\rho} \sum_{j=1}^{m+p} s_j^2 + \frac{1}{2\rho} \sum_{j=1}^{m+p} [2\lambda_j \rho Y_{c_j}(x) + 2s_j Y_{c_j}(x) + Y_{c_j}(x)^2] \\ &= Y_f(x) + \sum_{j=1}^{m+p} \lambda_j s_j + \frac{1}{2\rho} \sum_{j=1}^{m+p} s_j^2 + \frac{1}{2\rho} \sum_{j=1}^{m+p} [(\alpha_j + Y_{c_j}(x))^2 - \alpha_j^2], \end{aligned}$$

with $\alpha_j = \lambda_j \rho + s_j$. Now decompose the $Y(x, s)$ into a sum of three quantities:

$$\begin{aligned} Y(x, s) &= Y_f(x) + r(s) + \frac{1}{2\rho} W(x, s), \quad \text{with} \\ r(s) &= \sum_{j=1}^{m+p} \lambda_j s_j + \frac{1}{2\rho} \sum_{j=1}^{m+p} s_j^2 - \frac{1}{2\rho} \sum_{j=1}^{m+p} \alpha_j^2 \quad \text{and} \quad W(x, s) = \sum_{j=1}^{m+p} (\alpha_j + Y_{c_j}(x))^2. \end{aligned} \quad (11)$$

Using $Y_{c_j} \sim \mathcal{N}(\mu_{c_j}(x), \sigma_{c_j}^2(x))$, i.e., leveraging Gaussianity, W can be written as

$$W(x, s) = \sum_{j=1}^{m+p} \sigma_{c_j}^2(x) X_j(x, s), \quad \text{with } X_j(x, s) \sim \chi^2 \left(\text{dof}=1, \delta = \left(\frac{\mu_{c_j}(x) + \alpha_j}{\sigma_{c_j}(x)} \right)^2 \right). \quad (12)$$

The line above is the expression of a weighted sum of non-central chi-square (WSNC) variates. Each of the $m + p$ variates involves a unit degrees-of-freedom (dof) parameter, and a non-centrality parameter δ . A number of efficient methods exist for evaluating the density, distribution, and quantile functions of WSNC random variables (e.g., Davies, 1980; Duchesne and Lafaye de Micheaux, 2010). The R packages `CompQuadForm` (Lafaye de Micheaux, 2013) and `sadists` (Pav, 2015) provide library implementations. Details and code are provided in Appendix C.

Some constrained optimization problems involve an *a priori* known objective $f(x)$. In that case, referring back to (11), we are done: the distribution of $Y(x, s)$ is WSNC (as in (12)) shifted by a known quantity $f(x) + r(s)$, determined by particular choices of inputs x and slacks s . When $Y_f(x)$ is conditionally Gaussian we have that $\tilde{W}(x, s) = Y_f(x) + \frac{1}{2\rho}W(x, s)$ is the weighted sum of a Gaussian and WNCS variates, a problem that is again well-studied. The same methods cited above are easily augmented to handle that case—see Appendix C.

3.2 Slack-AL expected improvement

Evaluating EI at candidate (x, s) locations under the AL-composite involves working with $\text{EI}(x, s) = \mathbb{E}[(y_{\min}^n - Y(x, s)) \mathbb{I}_{\{Y(x, s) \leq y_{\min}^n\}}]$, given the current minimum y_{\min}^n of the AL over all n runs. When $f(x)$ is known, let $w_{\min}^n(x, s) = 2\rho(y_{\min}^n - f(x) - r(s))$ absorb all of the non-random quantities involved in the EI calculation. Then, with $D_W(\cdot; x, s)$ denoting the distribution of $W(x, s)$,

$$\begin{aligned} \text{EI}(x, s) &= \frac{1}{2\rho} \mathbb{E}[(w_{\min}^n(x, s) - W(x, s)) \mathbb{I}_{W(x, s) \leq w_{\min}^n(x, s)}] \\ &= \frac{1}{2\rho} \int_{-\infty}^{w_{\min}^n(x, s)} D_W(t; x, s) dt = \frac{1}{2\rho} \int_0^{w_{\min}^n(x, s)} D_W(t; x, s) dt \end{aligned} \quad (13)$$

if $w_{\min}^n(x, s) \geq 0$ and zero otherwise. That is, the EI boils down to integrating the distribution function of $W(x, s)$ between 0 (since W is positive) and $w_{\min}^n(x, s)$. This is a one-dimensional definite integral that is easy to approximate via quadrature; details are in Appendix C. Since $W(x, s)$ is quadratic in the $Y_c(x)$ values, it is often the case, especially for smaller ρ -values in later AL iterations, that $D_W(t; x, s)$ is zero over most of $[0, w_{\min}^n(x, s)]$, simplifying the numerical integration. However, this has deleterious impacts on search over (x, s) . We discuss a useful adaptation for that case in Appendix B.2.

When $f(x)$ is unknown and $Y_f(x)$ is conditionally normal, let $\tilde{w}_{\min}^n(s) = 2\rho(y_{\min}^n - r(s))$. Then,

$$\text{EI}(x, s) = \frac{1}{2\rho} \mathbb{E}[(\tilde{w}_{\min}^n(s) - \tilde{W}(x, s)) \mathbb{I}_{\tilde{W}(x, s) \leq \tilde{w}_{\min}^n(s)}] = \frac{1}{2\rho} \int_{-\infty}^{\tilde{w}_{\min}^n(s)} D_{\tilde{W}}(t; x, s) dt.$$

Here the lower bound of the definite integral cannot be zero since $Y_f(x)$ may be negative, and thus $\tilde{W}(x, s)$ may have non-zero distribution for negative t -values. This presents some challenges when it comes to numerical integration via quadrature, although many library functions allow indefinite bounds. We obtain better performance by supplying a conservative finite lower bound, for example three standard deviations in $Y_f(x)$, in units of the penalty (2ρ) , below zero: $6\rho\sigma_f(x)$. Detailed example implementations are provided Appendix C.

3.3 AL updates, optimal slacks, and other implementation notes

The new slack-AL method is completed by describing when the subproblem (9) is deemed to be “solved” (step 2 in Alg. 1), how λ and ρ updated (steps 3–4). We terminate the BO search sub-solver after a single iteration as this matches with the spirit of EI-based search, whose choice of next location can be shown to be optimal, in a certain sense, if it is the final point being selected (Bull, 2011). It also meshes well with an updating scheme analogous to that in steps 3–4: updating only when *no* actual improvement (in terms of constraint violation) is realized by that choice. For those updates, the analog for the mixed constraint setup is

- step 2: Let (x^k, s^k) approx. solve $\min_{x,s} \left\{ L_A(x, s; \lambda^{k-1}, \rho^{k-1}) : (x, s_{1:m}) \in \tilde{\mathcal{B}} \right\}$
- step 3: $\lambda_j^k = \lambda_j^{k-1} + \frac{1}{\rho^{k-1}}(c_j(x^k) + s_j^k)$, for $j = 1, \dots, m+p$
- step 4: If $c_{1:m}(x^k) \leq 0$ and $|c_{m+1:m+p}(x^k)| \leq \epsilon$, set $\rho^k = \rho^{k-1}$; else $\rho^k = \frac{1}{2}\rho^{k-1}$

Above, step 3 is the same as in Alg. 1 except without the “max”, and with slacks augmenting the constraint values. The “if” statement in step 4 checks for validity at x^k , deploying a threshold $\epsilon \geq 0$ on equality constraints; further discussion of the threshold ϵ is deferred to Section 4, where we discuss progress metrics under mixed constraints. If validity holds at (x^k, s^k) , the current AL iteration is deemed to have “made progress” and the penalty remains unchanged; otherwise it is doubled. An alternate formulation may check $|c_{1:m}(x^k) + s_{1:m}^k| \leq \epsilon$. We find that the version in step 4, above, is cleaner because it limits sensitivity to the choice of threshold ϵ . In Appendix B.1 recommend initial (λ^0, ρ^0) values which are analogous to the original, non-slack AL settings.

Optimal choice of slacks: The biggest difference between the original AL (3) and slack-AL (9) is that the latter requires searching over both x and s , whereas the former involves only x -values. In what follows we show that there are automatic choices for the s -values as a function of the corresponding x ’s, keeping the search space d -dimensional, rather than $d + m$.

For an observed $c_j(x)$ value, associated slack variables minimizing the AL (9) can be obtained analytically. Using the form of (11), observe that $\min_{s \in \mathbb{R}_+^m} y(x, s)$ is equivalent to $\min_{s \in \mathbb{R}_+^m} \sum_{j=1}^m 2\lambda_j \rho s_j + s_j^2 + 2s_j c_j(x)$. For fixed x , this is strictly convex in s . Therefore, its unconstrained minimum can only be its stationary point, which satisfies $0 = 2\lambda_j \rho + 2s_j^*(x) + 2c_j(x)$, for $j = 1, \dots, m$. Accounting for the nonnegativity constraint, we obtain the following optimal slack as a function of x :

$$s_j^*(x) = \max \{0, -\lambda_j \rho - c_j(x)\}, \quad j = 1, \dots, m. \quad (14)$$

Above we write s^* as a function of x to convey that x remains a “free” quantity in $y(x, s^*(x))$. Recall that slacks on equality constraints are zero, $s_k(x) = 0$, $k = m + 1, \dots, m + p$, for all x .

In the blackbox $c(x)$ setting, $y(x, s^*(x))$ is only directly accessible at the data locations x_i . At other x -values, however, the surrogates provide a useful approximation. When $Y_c(x)$ is (approximately) Gaussian it is straightforward to show that the optimal setting of the slack variables, solving $\min_{s \in \mathbb{R}_+^p} \mathbb{E}[Y(x, s)]$, are $s_j^*(x) = \max\{0, -\lambda_j \rho - \mu_{c_j}(x)\}$, i.e., the same as (14) with a prediction $\mu_{c_j}(x)$ for $Y_{c_j}(x)$ unknown $c_j(x)$ value. Again, slacks on the equality constraints are set to zero.

Other criteria can be used to choose slack variables. Instead of minimizing the mean of the composite, one could maximize the EI. In Appendix A we explain how this is of dubious practical value. Compared to the settings described above, searching over EI is both more computationally intensive and provides near identical results in practice.

Implementation notes: Code supporting all methods in this manuscript is provided in two open-source R packages: **laGP** (Gramacy, 2015) and **DiceOptim** (Ginsbourger et al., 2015), both on CRAN (R Development Core Team, 2004). Implementation details vary somewhat across those packages, due primarily to particulars of their surrogate modeling capability and how they search the EI surface. For example, **laGP** can accommodate a smaller initial design size because it learns fewer parameters (i.e., has fewer degrees of freedom). **DiceOptim** uses a multi-start search procedure for EI, whereas **laGP** deploys a random candidate grid, which may optionally be “finished” with an L-BFGS-B search. Nevertheless, their qualitative behavior exhibits strong similarity. Both packages also implement the original AL scheme (i.e., without slack variables) updated (8) for mixed constraints. Further details are provided in Appendix B.2.

4 Empirical comparison

Here we describe three test problems, each mixing challenging elements from traditional unconstrained blackbox optimization benchmarks, but in a constrained optimization format. We run our optimizers on these problems 100 times under random initializations. In the case of our GP surrogate comparators, this initialization involves choosing random space-filling designs. Our primary means of comparison is an averaged (over the 100 runs) measure of progress defined by the best valid value of the objective for increasing budgets (number of evaluations of the blackbox), n .

In the presence of equality constraints it is necessary to relax this definition somewhat, as the valid set may be of measure zero. In such cases we choose a tolerance $\epsilon \geq 0$ and declare a solution to be “valid” when inequality constraints are all valid, and when $|h_k(x)| \leq \epsilon$ for all $k = 1, \dots, p$. In our figures we choose $\epsilon = 10^{-2}$; however, the results are similar under stronger thresholds, with a higher variability over initializations. As finding a valid solution is, in itself, sometimes a difficult task, we additionally report the proportion of runs that find valid and optimal solutions as a function of budget, n , for problems with equality (and mixed) constraints.

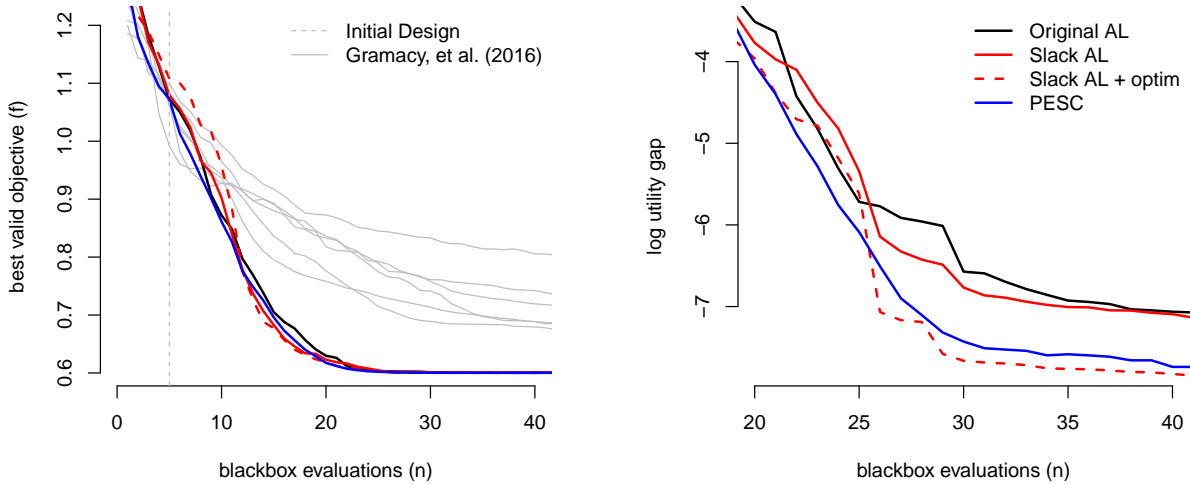


Figure 1: Results on the LSQ problem with initial designs of size $n = 10$. The *left* panel shows the best valid value of the objective over the first 40 evaluations, whereas the right shows the log *utility-gap* for the second 20 evaluations. The solid gray lines show comparators from Gramacy et al. (2016a).

4.1 An inequality constrained problem

We first revisit the “toy” problem from Gramacy et al. (2016a), having a 2-d input space limited to the unit cube, a (known) linear objective, with sinusoidal and quadratic inequality constraints (henceforth LSQ problem; see the Appendix D for details). Figure 1 shows progress over repeated solves with a maximum budget of 40 blackbox evaluations. The *left*-hand plot in Figure 1 tracks the average best valid value of the objective found over the iterations, using the progress metric described above. Random initial designs of size $n = 5$ were used, as indicated by the vertical-dashed gray line. The solid *gray* lines are extracted from a similar plot from Gramacy et al. (2016a), containing both AL-based comparators, and several from the derivative-free optimization and BO literatures. The details are omitted here. Our new ALBO comparators are shown in thicker colored lines; the solid black line is the original AL(BO)-EI comparator, under a revised (compared to (Gramacy et al., 2016a)) initialization and updating scheme. The two red lines are variations on the slack-AL algorithm under EI: with (dashed) and without (solid) L-BFGS-B optimizing EI acquisition at each iteration. Finally, the blue line is PESC Hernández-Lobato et al. (2015), using the Python library available at <https://github.com/HIPS/Spearmint/tree/PESC>. The take-home message from the plot is that all four new methods outperform those considered by the original ALBO paper Gramacy et al. (2016a).

Focusing on the new comparators only, observe that their progress is *nearly* statistically equivalent during the first 20 iterations. At iteration 10, for example, the ordering from best to worst is PESC (0.861), Slack-AL (0.902), Orig-AL (0.942), and Slack-AL+optim (0.957). Combining the 100 repetitions of each run, only the best (PESC) versus worst (Slack-AL+optim) is statistically significant in a one-sided t -test, with a p -value of 0.0011.

Apparently, aggressively optimizing the EI in the slack formulation in early iterations hurts more than it helps. The situation is more nuanced, however, for later iterations. At iteration 30, for example, the ordering is Slack-AL+optim (0.6002), PESC (0.6004), Slack-AL (0.6010), and Orig-AL (0.689). Although the numbers are more tightly grouped, only the first two (Slack-AL+optim and PESC), both leveraging L-BFGS-B subroutines, are statistically equivalent. In particular, both Slack-AL variants outperform the original AL with p -values less than $2.2e^{-16}$. This discrepancy is more easily visualized in the *left* of the figure with a so-called “utility-gap” log plot (Hernández-Lobato et al., 2015), which tracks the log difference between the theoretical best valid value and those found by search.

4.2 Mixed inequality and equality constrained problems

Next consider a problem in four input dimensions with a (known) linear objective and two constraints, one inequality and one equality. The inequality constraint is the so-called “Ackley” function in $d = 4$ input dimensions. The equality constraint follows the so-called “Hartman 4-dimensional function”. Appendix D provides a full mathematical specification. Figure 2 shows two views into progress on this problem. Since it involves mixed constraints,

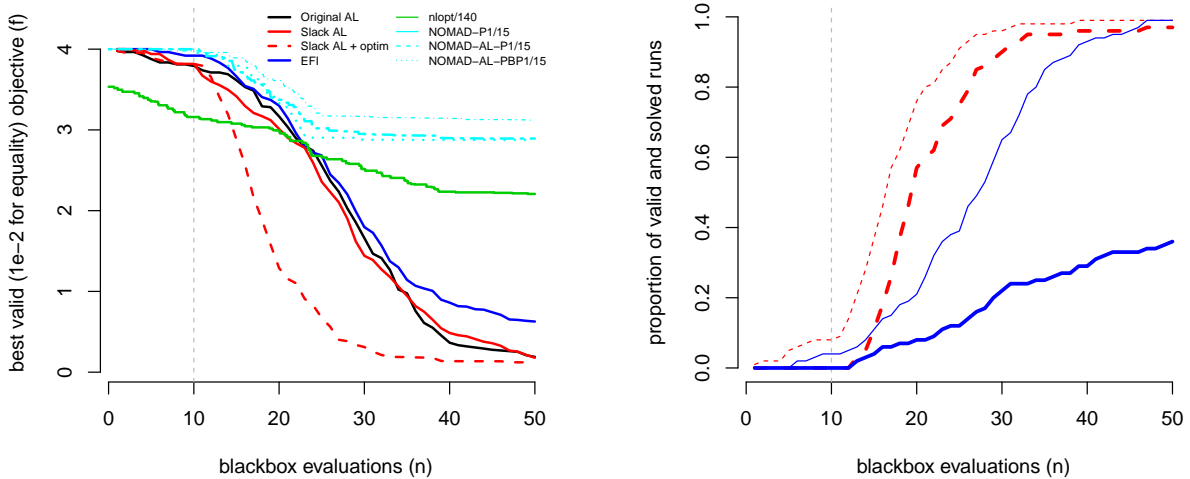


Figure 2: Results on the Linear-Hartman-Ackley mixed constraint problem. The *left* panel shows a progress comparison based on `laGP` code with initial designs of size $n = 10$. The x -scale has been divided by 140 for the `nlopt` comparator. A value of four indicates that no valid solution has been found. The *right* panel shows the proportion of valid (thin lines) and optimal (thick lines) solutions for the EFI and “Slack AL + optim” comparators.

comparators from the BO literature are scarce. Our EFI implementation deploys the $(-h, h)$ heuristic mentioned in the introduction. As representatives from the nonlinear optimization literature we include `nlopt` (Johnson, 2014) and three adapted `NOMAD` (Le Digabel, 2011) comparators, which are detailed in Appendix B.3. In the *left*-hand plot we can see that our new ALBO comparators are the clear winner, with an L-BFGS-B optimized EI search

under the slack-variable AL implementation performing exceptionally well. The **nlopt** and **NOMAD** comparators are particularly poor. We allowed those to run up to 7000 and 1000 iterations, respectively, and in the plot we scaled the x -axis (i.e., n) to put them on the same scale as the others. The *right*-hand plot provides a view into the distribution of two key aspects of performance over the MC repetitions. Observe that “Slack AL + optim” finds valid values quickly, and optimal values not much later. Our adapted EFI is particularly slow at converging to optimal (valid) solutions.

Our final problem involves two input dimensions, an *unknown* objective function (i.e., one that must be modeled with a GP), one inequality constraint and two equality constraints. The objective is a centered and re-scaled version of the “Goldstein–Price” function. The inequality constraint is the sinusoidal constraint from the LSQ problem [Section 4.1]. The first equality constraint is a centered “Branin” function, the second equality constraint is taken from Parr et al. (2012) (henceforth GBSP problem). Appendix D contains a full mathematical specification. Figure 3 shows our results on this problem. Observe (*left* panel)

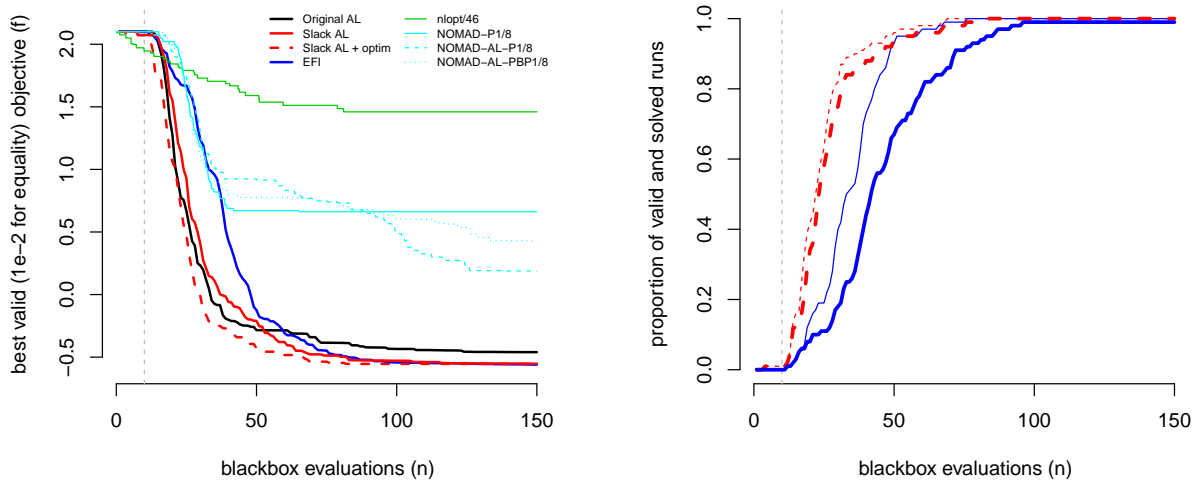


Figure 3: Results on the GBSP problem. See Figure 2 caption.

that the original ALBO comparator makes rapid progress at first, but dramatically slows for later iterations. The other ALBO comparators, including EFI, converge much more reliably, with the “Slack AL + optim” comparator leading in both stages (early progress and ultimate convergence). Again, **nlopt** and **NOMAD** are poor, however note that their relative comparison is reversed; again, we scaled the x -axis to view these on a similar scale as the others. The *right* panel shows the proportion of valid and optimal solutions for “Slack AL + optim” and EFI. Notice that the AL method finds an optimal solution almost as quickly as it finds a valid one—both substantially faster than EFI.

5 Discussion

The augmented Lagrangian (AL) is an established apparatus from the mathematical optimization literature, enabling unconstrained or bound-constrained optimizers to be deployed in settings with constraints. Recent work involving Bayesian optimization within the AL framework (ALBO) has shown great promise, especially toward obtaining global solutions under constraints. However, those methods were deficient in at least two respects. One is that only inequality constraints could be supported. Another was that evaluating the acquisition function, combining predictive mean and variance information via EI, required Monte Carlo approximation. In this paper we showed that both drawbacks could be addressed via a slack-variable reformulation of the AL. Our method supports inequality, equality, and mixed constraints, and to our knowledge this updated ALBO procedure is unique in the BO literature in its applicability to the general mixed constraints problem (1). We showed that the slack ALBO method outperforms modern alternatives in several challenging constrained optimization problems.

We conclude by remarking on a potential drawback: we have found in some cases that our slack variable approach is a double-edged sword, especially when the unknown slacks are chosen in a default manner, i.e., $s^*(x)$ as in Section 3.3. Those choices utilize the surrogate(s) of the constraints, particularly the posterior mean, which means that those surrogates are relied upon more heavily than in the original (non-slack) AL context. Consequently, if the problem at hand matches the surrogate modeling (i.e., Gaussian process) assumptions well, then by a leveraging argument we can expect the slack method to do better than the original. However, if the assumptions are a mismatch, then one can expect them to be worse. The motivating “Lockwood” problem from Gramacy et al. (2016a) had a constraint surface that was kinked at its boundary $\{x \in \mathbb{R}^d : c(x) = 0\}$, which obviously violates the smoothness assumptions of typical GP surrogates. We were therefore not surprised to find poorer performance in our slack method compared to the original.

Acknowledgments

We are grateful to Mickael Binois for comments on early drafts. RBG is grateful for partial support from National Science Foundation grant DMS-1521702. The work of SMW is supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research under Contract No. DE-AC02-06CH11357. The work of SLD is supported by the Natural Sciences and Engineering Research Council of Canada grant 418250.

A Optimal and default slack variables

This continues the discussion from Section 3.3, which suggests choosing slack variables by minimizing the slack-AL. One alternative is to maximize the EI. Recall that EI mixes data quantities (at the x_i) and predictive quantities at candidate locations (x, s) via the im-

provement $I(x, y) = \max(0, y_{\min}^n - Y(x, s))$, the former coming from $y_{\min}^n = \min(y_1, \dots, y_n)$, where $y_i = y(x_i, s)$. If one seeks improvement over the lowest possible y_{\min}^n , then the optimal s -values to pair with the x_i s are given in Eq. (14): use $y_i = y(x_i, s^*(x_i))$. Choosing slacks for candidate (x, s) locations is more challenging. Solving $\max_{s \in \mathbb{R}_+^m} \text{EI}(x, s)$ for a fixed x is equivalent to solving $\max_{s \in \mathbb{R}_+^m} \mathbb{E}\{Y(x, s) \mathbb{I}_{Y(x, s) \leq y_{\min}^n}\}$, for which a closed-form solution remains elusive. Numerical methods are an option; however, we have not discovered any advantage over the simpler $\min_{s \in \mathbb{R}_+^m} \mathbb{E}\{Y(x, s)\}$ -based settings described above. Empirically, we find that the two criteria either yield identical $s_j^*(x)$ -values, or ones that are nearly so.

As an illustration, we consider a 2-d problem with a linear objective and two inequality constraints (see the LSQ problem in Section D, Eq. (15), below). We focus here on a “static” situation, where a 9-point grid defines the set of observations. f_1 is treated as a known function and two GP models are fitted to c_1 and c_2 . The AL parameters are set to $\rho = 1/16$ and $\lambda = [1, 1]$. The *top-left* panel of Figure 4 shows the EI surface (computed on a 2-d grid), when $s_j^*(x) = \max(0, -\lambda_j \rho - \mu_j(x))$ is used for each grid element x . The shape of the contours is typical of EI, with a plateau of zero values and two local maxima. The remaining panels show the value of the EI metric for each of the three x -values shown in the *top-left* panel. These are computed on a 2-d grid in s space. Despite its complex formulation (13), on this example EI is apparently unimodal with respect to s_1 and s_2 . Moreover, the optimal values (computed from the grid used to plot the image) coincide (up to numerical error) with the associated s^* values, plotted as diamonds matching the color in the *top-left* plot.

B Implementation notes

B.1 Initializing the AL parameters in the slack setup

To initialize (λ^0, ρ^0) , we update the settings suggested in Gramacy et al. (2016b)—balancing the scales of objective and constraint in the AL on the initial design of size n_0 —to our new mixed constraint context. Let $v(x)$ be a logical vector of length $m + p$ recording the validity of x in a zero-slack setting. That is, let $v_j(x) = 1$ if $c_j(x) \leq 0$, for $j = 1, \dots, p$, and $v_j(x) = 1$ if $|c_j(x)| \leq \epsilon$, for $j = p + 1, \dots, m$; otherwise let $v_j(x) = 0$. Then, take

$$\rho^0 = \frac{\min_{i=1, \dots, n_0} \{\sum_{j=1}^{m+p} c_j(x_i)^2 : \exists j, v_j(x_i) = 0\}}{2 \min_{i=1, \dots, n_0} \{f(x_i) : \forall j, v_j(x_i) = 1\}},$$

and $\lambda^0 = 0$. The denominator above is not defined if the initial design has no valid values (i.e., if there is no x_i with $v_j(x_i) = 1$ for all j). When that happens we use the median of $f(x_i)$ in the denominator instead. On the other hand, if the initial design has no invalid values and hence the numerator is not defined, we use $\rho^0 = 1$.

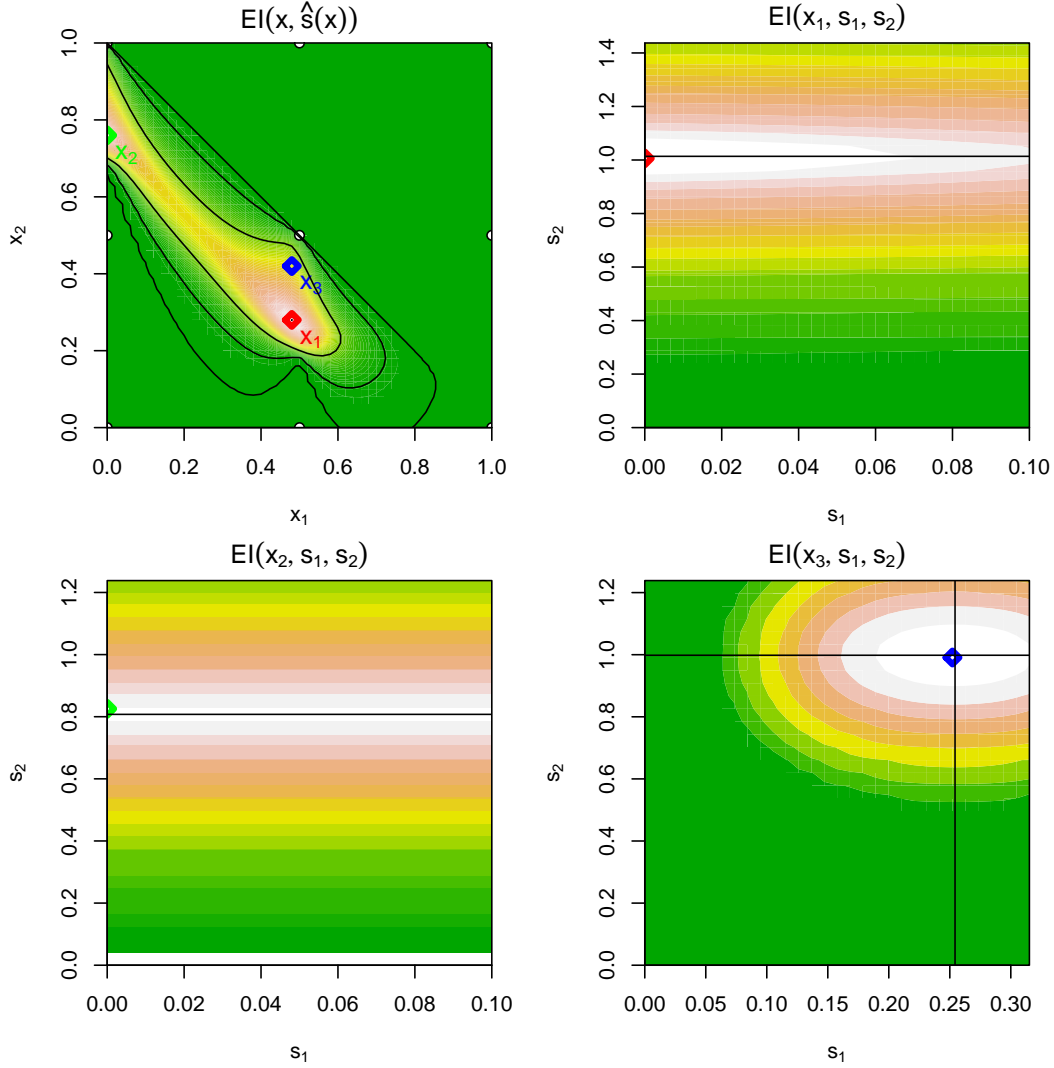


Figure 4: Illustration of the influence of the slack variables. The top left figure shows EI as a function of x , when the slack variables are chosen as \hat{s} . The three other figures show the EI metric as a function of the slack variables for three different x -values (at locations shown in the top left graph). The horizontal and vertical lines show the s^* -values, and the points show the optimal values.

B.2 Open source software

Code supporting all methods described here is provided in two open source R packages: **laGP** (Gramacy, 2015) and **DiceOptim** (Ginsbourger et al., 2015) packages, both on CRAN. Implementation details vary somewhat across those packages, due primarily to particulars of their surrogate modeling capability and how they search the EI surface. For example **laGP** can accommodate a smaller initial sample size because it learns fewer parameters (i.e., has fewer degrees of freedom). **DiceOptim** uses a multi-start search procedure for EI, whereas

laGP deploys a random candidate grid, which may optionally be “finished” with an L-BFGS-B search. Nevertheless, their qualitative behavior exhibits strong similarity. Both utilize library subroutines for dealing with weighted non-central chi-square deviates, as described in Appendix C below. Some notable similarities and differences are described below. The empirical work we present in Section 4 provides results from our **laGP** only. Those obtained from **DiceOptim** are similar; we encourage readers to consult the documentation from both packages for further illustration.²

Both initialize ρ and λ as described above in Section B.1 above; both update those parameters according to the modifications of steps 2–4 in Alg. 1; and both define x^k in those updates to be the value of x_i corresponding to the best $y_i = y(x_i, s^*(x_i))$ using $(\lambda^{k-1}, \rho^{k-1})$, for data values indexed by $i = 1, \dots, n$. By default, both utilize $(x, s^*(x))$ to evaluate EI at candidate x -locations, via the surrogate mean $\mu_c(x)$ values. The **laGP** implementation optionally allows randomization of $(x, S^*(x))$, where $S^*(x)$ is chosen via (14) with a draw $Y_{c_j}(x)$ in place of $c_j(x)$. Alternatively, **DiceKriging** can optimize jointly over (x, s) via its multi-start search scheme. Those two options are variations which allow one to test the “optimality” of $s^*(x)$ under EI calculations. We report here that such approaches have been found to be uniformly inferior to using $s^*(x)$, which is anyways a much simpler implementation choice.

Choosing $s^*(x)$, as opposed to random $S^*(x)$ above, renders the entire EI calculation for a candidate x value deterministic. In turn, that means that local (numerical) derivative-based optimization techniques, such as L-BFGS-B (Liu and Nocedal, 1989), can be used to optimize over EI settings. The **laGP** package has an optional “**optim**” setting, initializing local L-BFGS-B searches from the best random candidate point via. **DiceKriging** uses the *genoud* algorithm (GENetic Optimization Using Derivatives, Mebane Jr and Sekhon, 2011) to handle potentially multi-modal EI surfaces (Figure 4). Neither of these options is available for the original AL (without slacks), since evaluating the EI required Monte Carlo, effectively limiting the resolution of search. In later iterations, the original AL would fail to “drill” down into local troughs of the EI surface. In our empirical work [Section 4], we show that being able to optimize over the EI surface is crucial to obtaining good performance at later iterations. The *predictive entropy search* adaptation for constrained optimization (PESC; Hernández-Lobato et al., 2015) is similarly deterministic, and also utilizes L-BFGS-B to optimize over selections. We conjecture that the correspondingly suboptimal selections made by the original AL, and the poor initialization of updates of ρ pointed out by Picheny et al. (2016), explains the AL-versus-PESC outcome reported by Hernández-Lobato et al. (2015). Our revised experiments show a reversed (if extremely close) comparison.

Finally, we find that the following simple-yet-very-efficient numerical trick can be helpful when the EI surface is mostly zero, as often happens at later stages of optimization—a behavior that can be attributed to the quadratic nature of the AL, guaranteeing zero improvement in certain situations (Gramacy et al., 2016a). In the original **laGP** implementation

²At the time of writing, updated versions of **laGP** and **DiceOptim** are not “pushed” to CRAN; alpha testing versions are available from the authors upon request. Subsequent CRAN updates will include all slack AL enhancements made for this paper.

this drawback was addressed by “switching” to mean-based search (rather than EI). In the slack variable formulation, observe that zero EI may be realized when $w_{\min}^n < 0$, in which case the w_{\min}^n value itself (as it is less than zero) may stand in as a sensible replacement: capturing both mean and (deterministic aspects) of the penalty information. For example, when optimizing over EI in `DiceOptim` we find that this w_{\min}^n replacement helps the solver escape large plateaus of (otherwise) zero values.

B.3 Adapted NOMAD comparators

The three methods `NOMAD-P1`, `NOMAD-AL-P1`, and `NOMAD-AL-PB-P1` are based on the `NOMAD` software Le Digabel (2011) that implements the Mesh Adaptive Direct Search derivative-free algorithm Audet and Dennis, Jr. (2006) using the Progressive Barrier (PB) technique Audet and Dennis, Jr. (2009) for inequality constraints. All three methods begin with a first phase (P1) that focuses on obtaining a feasible solution, with the execution of `NOMAD` on the minimization of $\sum_{k=1}^p h_k^2(x)$ subject to $g_j(x) \leq 0, j = 1, \dots, m$. Then, `NOMAD-P1` follows with an ordinary `NOMAD` run where equalities are transformed into inequalities of the form $|h_k(x)| \leq 0, k = 1, \dots, p$. `NOMAD-AL-P1` is a straightforward extension of the augmented Lagrangian method of Gramacy et al. (2016a), with the inclusion of equality constraints in the Lagrangian function, while the unconstrained subproblem is handled by `NOMAD`. `NOMAD-AL-PB-P1` is a hybrid version where equalities are still in the Lagrangian while inequalities are transferred into the subproblem where `NOMAD` treats them with the PB. Note that the non-use of the first phase P1 was also tested, but it did not work as well as the others so it was not included in order to save space.

C Code for calculations with WSNC RVs

Here we provide R code for calculating EI and related quantities using the `sadists` (Pav, 2015) and `CompQuadForm` (Lafaye de Micheaux, 2013) libraries, in particular their WNCS distribution functions (e.g., Davies, 1980; Duchesne and Lafaye de Micheaux, 2010).

Let `mz` and `sz` denote the vectorized means and variances of the normal Z_i quantity in (12), that is `mx[j] = $\mu_{c_j}(x) + \alpha_j$` . Below, `rho` contains the AL parameter ρ , and we presume $f(x)$, $r(s)$, and y_{\min}^n quantities have been pre-calculated and stored in variables of the same name. Then, the relevant quantities for calculating EI in the case of a known $f(x)$ can be computed as:

```
R> ncp <- (mz / sz)^2
R> wmin <- 2*rho*(ymin - fx - gs)
```

Using `sadists`, EI may be computed as follows using simple “quadrature”.

```
R> library(sadists)
R> m <- length(mz)
R> lt <- 1000
```

```

R> t <- seq(0, wmin, length=lt)
R> df <- pow <- rep(1, mz)
R> ncp <- (mx / sx)^2
R> EI <- (sum(psumchisqpow(q=t, sz^2, df, ncp, pow))*wmin/(lt-1)) / (2*rho)

```

Alternately, via the `integrate` function built-into R:

```

R> EIgrand <- function(t) { psumchisqpow(q=t, wts, df, ncp, pow) }
R> EI <- integrate(EIgrand, 0, wmin)$value / (2*rho)

```

The code using `CompQuadForm` is similar. Greater care is required to code the EI integrand, as the main function (`davies`) is not vectorized. Also, we have found that NaN is often erroneously returned when the integrand is actually zero.

```

library(CompQuadForm)
R> EIgrand <- function(x) {
+   p <- rep(NA, length(x))
+   for(1 in 1:length(c))
+     p[i] <- 1 - davies(q=x[i], lambda=sz^2, delta=ncp)$Qq
+   p[!is.finite(p)] <- 0
+   return(p)
+ }
R> EI <- (sum(EIgrand(t))*wmin/(lt-1)) / (2*rho)
R> EI <- integrate(EIgrand, 0, wmin)$value / (2*rho) ## alternately

```

Although this may at first seem more cumbersome, it is actually fairly easy to vectorize `davies`, and at the same time replace NaN values with zero in the underlying C implementation. The result is a method which is much faster than the `sadists` alternative, and has the added bonus of being callable from the C functions inside our `laGP` implementation. For more details on this re-implementation of `davies` in our setup, please see the source for the `laGP` and/or `DiceOptim` package(s).

Some slight modification is required for the unknown objective case. Let `mf` and `sf` denote $\mu_f(x)$ and $\sigma_f(x)$ in the code below. We have concluded that only the methods from `CompQuadForm` are applicable in this case.

```

R> alpha <- 2*rho*lambda + 2*s
R> wmin <- 2*rho*ymin - sum(s^2) - s %%% t(lambda) + sum(alpha^2/4)
t <- seq(-6*rho*sf, wmin, length=lt)
R> EIgrand <- function(x) {
+   p <- rep(NA, length(x))
+   madj <- 2*rho*mf
+   sadj <- 2*rho*sf
+   for(1 in 1:length(c))
+     p[i] <- davies(q=x[i]-madj, lambda=sz^2, delta=ncp, sigma=sadj)$Qq
+ }

```

```

+     p <- 1-p
+     p[!is.finite(p)] <- 0
+     return(p)
}
R> EI <- (sum(EIgrand(t))*(wmin+6*rho*s)/(lt-1)) / (2*rho)
R> EI <- integrate(EIgrand, -6*rho*sf, wmin)$value / (2*rho) ## alternately

```

As when $f(x)$ is known, vectorizing the `davies` routine leads to dramatic speedups.

D Test problems

Below we detail the components of the three test problems explored in Section 4. They combine classic benchmarks from the Bayesian optimization and computer (surrogate) modeling literature. In total there are two objective functions and six constraints.

Objective functions: f_1 is a simple linear objective, treated as known, and f_2 is the *Goldstein-Price* function (rescaled and centered) Dixon and Szego (1978); Molga and Smutnicki (2005); Picheny et al. (2013). Further description, as well as R and MATLAB implementations, is provided by Bingham et al. (2014).

$$\begin{aligned}
f_1(x) &= \sum_{i=1}^d x_i \\
f_2(x) &= \frac{\log[(1+a)(30+b)] - 8.69}{2.43}, \quad \text{with} \\
a &= (4x_1 + 4x_2 - 3)^2 \times \\
&\quad [75 - 56(x_1 + x_2) + 3(4x_1 - 2)^2 + 6(4x_1 - 2)(4x_2 - 2) + 3(4x_2 - 2)^2] \\
b &= (8x_1 - 12x_2 + 2)^2 \times \\
&\quad [-14 - 128x_1 + 12(4x_1 - 2)^2 + 192x_2 - 36(4x_1 - 2)(4x_2 - 2) + 27(4x_2 - 2)^2]
\end{aligned}$$

Constraint functions: c_1 and c_2 are the “toy” constraints from Gramacy et al. (2016a); c_3 is the “Branin” function (centered) “Branin” function (Dixon and Szego, 1978; Forrester et al., 2008; Molga and Smutnicki, 2005; Picheny et al., 2013); c_4 is taken from Parr et al. (2012). c_5 is the “Ackley” function (centered) (for details see Adorio and Diliman, 2013; Molga and Smutnicki, 2005; Bäck, 1996); c_6 the “Hartman” function (centered, rescaled) (see Dixon and Szego, 1978; Picheny et al., 2013, for details). Again Surjanovic and Bingham (2014) provides a convenient one-stop reference containing R and MATLAB code and visualizations.

$$c_1(x) = 0.5 \sin(2\pi(x_1^2 - 2x_2)) + x_1 + 2x_2 - 1.5$$

$$c_2(x) = -x_1^2 - x_2^2 + 1.5$$

$$c_3(x) = 15 - \left(15x_2 - \frac{5}{4\pi^2}(15x_1 - 5)^2 + \frac{5}{\pi}(15x_1 - 5) - 6\right)^2 - 10\left(1 - \frac{1}{8\pi}\right)\cos(15x_1 - 5)$$

$$c_4(x) = 4 - \left(4 - 2.1(2x_1 - 1)^2 + \frac{(2x_1 - 1)^4}{3}\right)(2x_1 - 1)^2 - (2x_1 - 1)(2x_2 - 1) \\ - 16(x_2^2 - x_2)(2x_2 - 1)^2 - 3\sin[12(1 - x_1)] - 3\sin[12(1 - x_2)]$$

$$c_5(x) = 3 + 20 \exp\left(-0.2\sqrt{\frac{1}{4}\sum_{i=1}^4(3x_i - 1)^2}\right) + \exp\left(\frac{1}{4}\sum_{i=1}^4\cos(2\pi(3x_i - 1))\right) - 20 - \exp(1)$$

$$c_6(x) = \frac{1}{0.8387} \left[-1.1 + \sum_{i=1}^4 C_i \exp\left(-\sum_{j=1}^4 a_{ji}(x_j - p_{ji})^2\right) \right],$$

with:

$$\mathbf{C} = \begin{bmatrix} 1.0 \\ 1.2 \\ 3.0 \\ 3.2 \end{bmatrix}, \quad \mathbf{a} = \begin{bmatrix} 10.00 & 0.05 & 3.00 & 17.00 \\ 3.00 & 10.00 & 3.50 & 8.00 \\ 17.00 & 17.00 & 1.70 & 0.05 \\ 3.50 & 0.10 & 10.00 & 10.00 \end{bmatrix}, \quad \mathbf{p} = \begin{bmatrix} 0.131 & 0.232 & 0.234 & 0.404 \\ 0.169 & 0.413 & 0.145 & 0.882 \\ 0.556 & 0.830 & 0.352 & 0.873 \\ 0.012 & 0.373 & 0.288 & 0.574 \end{bmatrix}.$$

Problems:

$$(\text{LSQ}) \quad \min_{x \in [0,1]^2} f_1(x) \quad \text{s.t.} \quad c_1(x) \leq 0, \quad c_2(x) \leq 0 \quad (15)$$

$$(\text{GSBP}) \quad \min_{x \in [0,1]^2} f_2(x) \quad \text{s.t.} \quad c_1(x) \leq 0, \quad c_2(x) = 0, \quad c_3(x) = 0 \quad (16)$$

$$(\text{LAH}) \quad \min_{x \in [0,1]^4} f_1(x) \quad \text{s.t.} \quad c_5(x) \leq 0, \quad c_6(x) = 0 \quad (17)$$

Figure 5 shows the two dimensional problems.

References

- Adorio, E. and Diliman, U. (2013). “MVF: Multivariate Test Functions Library in C for Unconstrained Global Optimization.” Retrieved June 2013, from <http://www.geocities.ws/eadorio/mvf.pdf>.
- Audet, C. and Dennis, Jr., J. (2006). “Mesh Adaptive Direct Search Algorithms for Constrained Optimization.” *SIAM Journal on Optimization*, 17, 1, 188–217.
- (2009). “A Progressive Barrier for Derivative-Free Nonlinear Programming.” *SIAM Journal on Optimization*, 20, 1, 445–472.

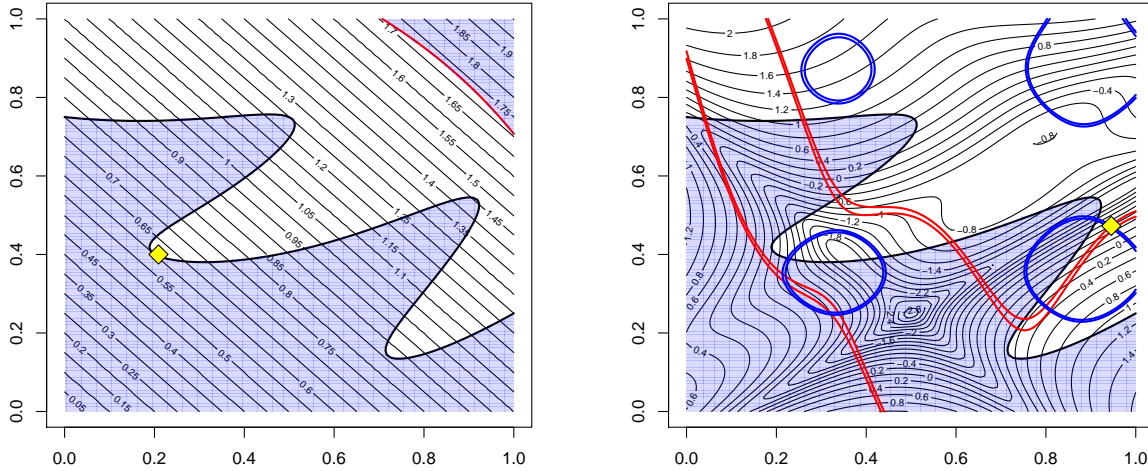


Figure 5: LSQ (left) and GSBP (right) problems. The contour lines show the objective functions, the bold lines the constraints, and the squares the solutions of the problems. Equality constraints are shown with two lines to represent the tolerance, the infeasible space defined by the inequality constraints are shaded.

Bäck, T. (1996). *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press.

Bertsekas (1982). *Constrained Optimization and Lagrange Multiplier Methods*. New York, NY: Academic Press.

Bingham, D., Ranjan, P., and Welch, W. (2014). “Sequential design of computer experiments for optimization, estimating contours, and related objectives.” In *Statistics in Action: A Canadian Outlook*, ed. J. Lawless, 109–124. Chapman & Hall.

Booker, A. J., Dennis, Jr., J. E., Frank, P. D., Serafani, D. B., Torczon, V., and Trosset, M. W. (1999). “A Rigorous Framework for Optimization of Expensive Functions by Surrogates.” *Structural Optimization*, 17, 1–13.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model Building and Response Surfaces*. Oxford: Wiley.

Boyle, P. (2007). “Gaussian Processes for Regression and Optimization.” Ph.D. thesis, Victoria University of Wellington.

Brochu, E., Cora, V. M., and de Freitas, N. (2010). “A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning.” Tech. rep., University of British Columbia. ArXiv:1012.2599v1.

Bull, A. D. (2011). “Convergence Rates of Efficient Global Optimization Algorithms.” *Journal of Machine Learning Research*, 12, 2879–2904.

- Davies, R. (1980). “Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29, 3, 323–333.
- Dixon, L. and Szego, G. (1978). “The global optimization problem: an introduction.” *Towards global optimization*, 2, 1–15.
- Duchesne, P. and Lafaye de Micheaux, P. (2010). “Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods.” *Computational Statistics and Data Analysis*, 54, 858–862.
- Forrester, A., Sobester, A., and Keane, A. (2008). *Engineering design via surrogate modelling: a practical guide*. Wiley.
- Gardner, J. R., Kusner, M. J., Xu, Z., Weinberger, K. W., and Cunningham, J. P. (2014). “Bayesian Optimization with Inequality Constraints.” In *Proceedings of the 31st International Conference on Machine Learning*, vol. 32. JMLR, W&CP.
- Gelbart, M. A., Snoek, J., and Adams, R. P. (2014). “Bayesian optimization with unknown constraints.” In *Uncertainty in Artificial Intelligence (UAI)*.
- Ginsbourger, D., Picheny, V., Roustant, O., with contributions by C. Chevalier, Marmin, S., and Wagner, T. (2015). *DiceOptim: Kriging-Based Optimization for Computer Experiments*. R package version 1.5.
- Gramacy, R. (2015). “1aGP: Large-Scale Spatial Modeling via Local Approximate Gaussian Processes in R.” *Journal of Statistical Software*, to appear. Available as a vignette in the 1aGP package.
- Gramacy, R., Gray, G., Le Digabel, S., Lee, H., Ranjan, P., Wells, G., and Wild, S. (2016a). “Modeling an Augmented Lagrangian for Blackbox Constrained Optimization.” *Technometrics*, 58, 1–11.
- (2016b). “Rejoinder to Modeling an Augmented Lagrangian for Blackbox Constrained Optimization.” *Technometrics*, 58, 1, 26–29.
- Gramacy, R. B. and Lee, H. K. H. (2011). “Optimization under Unknown Constraints.” In *Bayesian Statistics 9*, eds. J. Bernardo, S. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 229–256. Oxford University Press.
- Hernández-Lobato, J. M., Gelbart, M. A., Hoffman, M. W., Adams, R. P., and Ghahramani, Z. (2015). “Predictive Entropy Search for Bayesian Optimization with Unknown Constraints.” In *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37. JMLR, W&CP.
- Johnson, S. G. (2014). “The NLOpt nonlinear-optimization package.” Via the R package `nloptr`.

- Jones, D. R., Schonlau, M., and Welch, W. J. (1998). “Efficient Global Optimization of Expensive Black Box Functions.” *J. of Global Optimization*, 13, 455–492.
- Lafaye de Micheaux, P. (2013). *CompQuadForm: Distribution function of quadratic forms in normal variables*. R package version 1.4-1.
- Le Digabel, S. (2011). “Algorithm 909: NOMAD: Nonlinear Optimization with the MADS algorithm.” *ACM Transactions on Mathematical Software*, 37, 4, 44:1–44:15.
- Liu, D. C. and Nocedal, J. (1989). “On the limited memory BFGS method for large scale optimization.” *Mathematical Programming*, 45, 1-3, 503–528.
- Mebane Jr, W. R. and Sekhon, J. S. (2011). “Genetic optimization using derivatives: the *rgenoud* package for R.” *Journal of Statistical Software*, 42, 11, 1–26.
- Mockus, J. (1989). *Bayesian Approach to Global Optimization: Theory and Applications*. Springer.
- (1994). “Application of Bayesian Approach to Numerical Methods of Global and Stochastic Optimization.” *J. of Global Optimization*, 4, 347–365.
- Mockus, J., Tiesis, V., and Zilinskas, A. (1978). “The Application of Bayesian Methods for Seeking the Extremum.” *Towards Global Optimization*, 2, 117-129, 2.
- Molga, M. and Smutnicki, C. (2005). “Test functions for optimization needs.” Retrieved June 2013, from <http://www.zsd.ict.pwr.wroc.pl/files/docs/functions.pdf>.
- Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. 2nd ed. Springer.
- Parr, J., Keane, A., Forrester, A., and Holden, C. (2012). “Infill sampling criteria for surrogate-based optimization with constraint handling.” *Engineering Optimization*, 44, 1147–1166.
- Pav, S. E. (2015). *sadists: Some Additional Distributions*. R package version 0.2.1.
- Picheny, V. (2014). “A stepwise uncertainty reduction approach to constrained global optimization.” In *Proceedings of the 7th International Conference on Artificial Intelligence and Statistics*, vol. 33, 787–795. JMLR W&CP.
- Picheny, V., Ginsbourger, D., and Krityakierne, T. (2016). “Comment: Some Enhancements Over the Augmented Lagrangian Approach.” *Technometrics*, 58, 1, 17–21.
- Picheny, V., Wagner, T., and Ginsbourger, D. (2013). “A benchmark of kriging-based infill criteria for noisy optimization.” *Structural and Multidisciplinary Optimization*, 48, 3, 607–626.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

- Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. New York: Springer-Verlag.
- Sasena, M. J. (2002). “Flexibility and Efficiency Enhancement for Constrained Global Design Optimization with Kriging Approximations.” Ph.D. thesis, University of Michigan.
- Schonlau, M., Welch, W. J., and Jones, D. R. (1998). “Global versus local search in constrained optimization of computer models.” *Lecture Notes-Monograph Series*, 11–25.
- R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Aus. ISBN 3-900051-00-3.
- Snoek, J., Larochelle, H., Snoek, J., and Adams, R. P. (2012). “Bayesian optimization of machine learning algorithms.” In *Neural Information Processing Systems (NIPS)*.
- Surjanovic, S. and Bingham, D. (2014). “Virtual Library of Simulation Experiments: Test Functions and Datasets.” Retrieved December 4, 2014, from <http://www.sfu.ca/~ssurjano>.
- Williams, B. J., Santner, T. J., Notz, W. I., and Lehman, J. S. (2010). “Sequential Design of Computer Experiments for Constrained Optimization.” In *Statistical Modeling and Regression Structures*, eds. T. Kneib and G. Tutz, 449–472. Springer-Verlag.